Delivering the IMPACT project Polish Ground-Truth texts with Poliqarp for DjVu

Janusz S. Bień

17 May 2012

1 Introduction

The Ground-Truth text files has been made available by Poznań Supercomputing and Networking Center Digital Libraries Team at

`http://dl.psnc.pl/activities/projekty/impact/results`

on the liberal Creative Commons Attribution 3.0 Unported License

`http://creativecommons.org/licenses/by/3.0/`[1].

However in this form they are of little use for researchers, such as historical linguists, interested in their text content. Therefore the data has been imported to the search engine Poliqarp for DjVu developed in the framework of the Ministry of Science and Higher Education's grant no. N N519 384036 (`https://bitbucket.org/jsbien/ndt`).

The texts has been subject to minor corrections made by Janusz S. Bień and described in the note *Changes to the IMPACT project Polish Ground-Truth texts* (`http://bc.klf.uw.edu.pl/288/`). They are now presented in the form of two corpora using slightly different encoding:

- so called one-dimensional transcription `http://poliqarp.wbl.klf.uw.edu.pl/pl/IMPACT_GT_1/`,
- so called two-dimensional transcription `http://poliqarp.wbl.klf.uw.edu.pl/pl/IMPACT_GT_2/`.

In the one-dimensional transcription the hyphenated words are reconstructed while in the two-dimensional transcription the hyphenation is preserved in its original form. Additionaly in every corpus the texts are available in the two type of transcription:

- facsimile (or strict diplomatic) transcription, preserving a lot of typographical details;

[1]The Polish language version of the page identifies the license as *Licencja Creative Commons Creative Commons Uznanie autorstwa 3.0 Polska* (`http://creativecommons.org/licenses/by/3.0/pl/`).

- so called textel transcription.

The corpora has been created by Krzysztof Szafran using the following free tools:

- Poliqarp (`http://poliqarp.sourceforge.net/`,
- Poliqarp for DjVu extension by Jakub Wilk (`https://bitbucket.org/jwilk/marasca-wbl`), includes the converter of the PAGE XML format to hOCR,
- a transcription tool designed by Janusz S. Bień and implemented by Tomasz Olejniczak (`https://bitbucket.org/jsbien/pol`),
- various auxiliary scripts by Tomasz Olejniczak (available on request),
- a script by Mirosław Michalski for augmenting PAGE XML files with line coordinates (available on request).

The character histograms has been made with the unihistext program implemented by Piotr Findeisen (`https://bitbucket.org/jsbien/unihistext`).

2 The content

The titles of the works are given exactly in the form in which they occur in the corpus metadata. When appropriate, at the beginning of the item description the abbreviation is given which is used by the Late Middle Polish dictionary (*Słownik języka polskiego XVII i 1. połowy XVIII wieku*, `http://sxvii.pl/`).

2.1 Books

The list is ordered according to the number of pages.

- ChmielAtenyNW. Benedykt Chmielowski *Nowe Ateny*, 3027 pages.
  - `Nowe Ateny albo Akademia wszelkiey scyencyi pełna, na różne tytuły iak na classes podzielona, mądrym dla memoryału, idiotom dla nauki, politykom dla praktyki, melancholikom dla rozrywki erygowana ... / przez Xiędza Benedykta Chmielowskiego ... . Część 1.`[2]. `http://www.wbc.poznan.pl/publication/3735`, publication date 1756 (drugie wydanie), 844 pages.
  - `Nowe Ateny, albo Akademia wszelkiey scyencyi pełna, na różne tytuły iak na classes podzielona, mądrym dla memoryału, idiotom dla nauki, politykom dla praktyki, melancholikom dla rozrywki erygowana ... / przez Xiędza Benedykta Chmielowskiego ... . Część 2.`. `http://www.wbc.poznan.pl/publication/3736`, publication date 1746, 810 pages.

[2]The pages 373–378 are missing in the IMPACT data.

  - Nowe Ateny, albo Akademia wszelkiey scyencyi pełna, na różne tytuły iak na classes podzielona, mądrym dla memoryału, idiotom dla nauki, politykom dla praktyki, melancholikom dla rozrywki erygowana ... . Część 3 albo Supplement.. http://www.wbc.poznan.pl/publication/3754[3], publication date 1754, 741 pages.
  - Nowe Ateny, albo Akademia wszelkiey scyencyi pełna, na różne tytuły iak na classes podzielona, mądrym dla memoryału, idiotom dla nauki, politykom dla praktyki, melancholikom dla rozrywki erygowana ... . Część 4, a drugi Supplement.. http://www.wbc.poznan.pl/publication/3737, publication date 1756, 632 pages.

- DrużZbiór. Zbiór rytmów duchownych Panegirycznych Moralnych i Swiatowych [...] Elżbiety z Kowalskich Druzbackiey [...] Zebrany y do druku podany przez J. Z. R. K. O. W. etc. [Załuskiego Józefa Andrzeja]. http://www.wbc.poznan.pl/publication/13950, publication date 1752, 566 pages.
- SykstCiepl. Erazm Sixtus O cieplicach we Skle Ksiąg Troie. Przez Erazma Syxta Philozophiey y Medicyny Doktora Napisanych…. http://dlibra.bibliotekaelblaska.pl/publication/6186[4], rok publikacji 1617, 242 pages.
- HaurEk. Jakub Haur Oekonomika ziemianska generalna Punktámi Pártikulárnemi, Interrogatoryámi Gospodárskiemi, Praktyką Mieśięczną, Modelluszámi abo Tabułámi Arithmetycznemi obiaśniona. Pánom Dźiedzicznym, Arendarzom, Oekonomom, Attendętom, Urzędnikom, [...].. http://www.dbc.wroc.pl/publication/1459, publication date 1675, 195 pages.
- GrodDysk. Jan Grodwanger Discurs o cenie pieniedzy teraznieyszey y o niektorych skutkach iey… http://dlibra.bibliotekaelblaska.pl/publication/6254, rok publikacji 1632, 64 pages.

2.2 News pamphlets

All the pamphlets come from the *Digital Library of Polish and Poland-Related News Pamphlets* (http://cbdu.id.uw.edu.pl/). The list is ordered according to the publication date.

- Żałosne opisanie upadku króla hiszpańskiego na morzu i na lądzie. http://cbdu.id.uw.edu.pl/2240/, publication date 1589.

[3]The digital copy in Digital Library of Wielkopolska is missing some pages at the beginning.
[4]In the library the pages 138 and 139 are duplicated.

- Nowiny z Rakuz o monstrancji luterskiej. http://cbdu.id.uw.edu.pl/2270/, publication date 1590.
- Powodzenia niebezpiecznego ale szczęśliwego wojska j. k. m. w Multanach opisanie. http://cbdu.id.uw.edu.pl/2800/, publication date 1601.
- List o oblężeniu zamku Dyjamenckiego w Inflantach do Krzysztofa Moniwida Dorohostajskiego, dnia 22 października 1605 pisany. http://cbdu.id.uw.edu.pl/2910/, publication date 1605.
- Nowiny z Inflant o porażce, która się stała nad Karolem, księciem Sudermańskim przez Jana Karola Chodkiewicza dnia 27 września 1605. http://cbdu.id.uw.edu.pl/2920, publication date 1605.
- Diariusz wiadomości od wyjazdu króla z Wilna do Smoleńska. http://cbdu.id.uw.edu.pl/3040/, publication date 1610.
- Ceremonie i porządek w koronowaniu Marii de Medici, królowej francuskiej i nawarskiej 13 maja 1610. http://cbdu.id.uw.edu.pl/3050/, publication date 1610.
- Szturm pocieszny smoleński, który był odprawiony szczęśliwie 13 czerwca 1611. http://cbdu.id.uw.edu.pl/3110/, publication date 1611.
- Pasja żołnierzy obojga narodów w stolicy moskiewskiej krótko opisana. http://cbdu.id.uw.edu.pl/3180/, publication date 1613.
- Nowiny z Torunia o zabronieniu przez heretyków nabożeństwa i procesji katolickich. http://cbdu.id.uw.edu.pl/3220/, publication date 1614.
- Nowe nowiny z Czech, Tatar i Węgier, przy tym rewokacja księcia czeskiego i jak radę cesarską z zamku oknem wyrzucali i innego króla sobie obrali. http://cbdu.id.uw.edu.pl/3390/, publication date 1620.
- Chorągiew Sarmacka w Wołoszech, to jest pospolite ruszenie i szczęśliwy powrót Polaków z Wołoch w roku 1621. http://cbdu.id.uw.edu.pl/3450/, publication date 1621.
- Poseł z Wołoch z obozu polskiego. 1621. http://cbdu.id.uw.edu.pl/3460/, publication date 1621.
- Relacja prawdziwa o wejściu wojska polskiego do Wołoch i o potrzebie jego z pogaństwem we wrześniu i październiku 1620. http://cbdu.id.uw.edu.pl/3510/, publication date 1621.

- Adwersaria, albo terminata sprawy wojennej, która się toczyła w wołoskiej ziemi z tureckim cesarzem. http://cbdu.id.uw.edu.pl/3540/, publication date 1621.
- Wieść z Moskwy prawdziwa krótkim rymem wyprowadzona. http://cbdu.id.uw.edu.pl/4100/, publication date 1634.
- Nowiny z Moskwy albo wota z traktatów i konsulty panów radnych ziemi moskiewskiej, które carowi swemu podawali. http://cbdu.id.uw.edu.pl/4250/, publication date 1634.
- Wyprawa i wyjazd sułtana Amurata, cesarza tureckiego, na wojnę do Korony Polskiej. http://cbdu.id.uw.edu.pl/4340/, publication date 1634.
- Prawdziwa relacja i opisanie straszliwego trzęsienia ziemi 27 marca roku 1638 w Kalabrii. http://cbdu.id.uw.edu.pl/4470/, publication date 1638.
- Relacja koronacji cudownego obrazu Najświętszej Marii Panny na Górze Różańcowej [w Podkamieniu]. http://cbdu.id.uw.edu.pl/14410/, publication date 1728.
- Relacja chwalebnej ekspedycji Jana Kazimierza, króla polskiego i szwedzkiego. http://cbdu.id.uw.edu.pl/4880/, publication date 1650.
- Relacja spraw gdańskich na sejmie walnym warszawskim roku 1570. http://cbdu.id.uw.edu.pl/710/, publication date 1570.
- Sławna wiktoria nad Turkami od wojsk koronnych i Wielkiego Księstwa Litewskiego pod Chocimiem otrzymana. http://cbdu.id.uw.edu.pl/9230/, publication date 1674.
- SŁAWNA VICTORIA, NAD TVRKAMI. OD WOYSK KORONNYCH: Y WIELKIEGO XIĘSTWA LITEWSKIEGO: Pod CHOCIMEM OTRZYMANA. W dźień SWIĘTEGO MARCINA w Roku 1673.. http://cbdu.id.uw.edu.pl/9240/, publication date 1674.

3 Encoding

The texts are encoded in Unicode with some characters from the Private Use Area. With one exception the PUA characters are used according to the recommendations of the Medieval Unicode Font Initiative (http://www.mufi.info/), so for proper display a MUFI-compatible font, e.g. Junicode (http://junicode.sourceforge.net/), is required.

The exception concerns the character LATIN SMALL LIGATURE LONG S L WITH STROKE, for which the code point U+F51E is used (which has been assigned

to it in the IMPACT project). There is now no publicly available font supporting this character. Fortunately the character is used only in the facsimile transcription.

The histogram of characters occuring in the texts is enclosed in the appendix.

4 Text segmentation

Texts are segmented into words according to the Unicode Standard Annex #29 *Unicode Text Segmentation* (`http://unicode.org/reports/tr29/`), with the following *ad hoc* exceptions:

- all PUA characters are treated as letters,
- `REPLACEMENT CHARACTER` (`FFFD`, used for unredable characters) is treated as a letter,
- `NON-BREAK SPACE`, which in the corpus preceds only `COMBINING LATIN SMALL LETTER O`, is also treated as a letter.

In the one-dimensional transition the characters `HYPHEN-MINUS` (`002D`), `NON-BREAKING HYPHEN` (`2011`) and `DOUBLE OBLIQUE HYPHEN` (`2E17`) at the end of a line are removed together with the `LINE FEED (LF)` (`000A`) character.

Dropped capitals, which in the original PAGE format are separated from the rest of the word, in the trancriptions are just a part of the respective word.

5 Transcription

Describing the principles of the facsimile transcription is outside the scope of the present note.

The primary purpose of the so called textel[5] transcription is to facilitate searching by replacing the ligatures with the appropriate sequence of characters, e.g. `LATIN SMALL LIGATURE LONG S L WITH STROKE` is replaced by `LATIN SMALL LETTER LONG S` and `LATIN SMALL LETTER L WITH STROKE`.

Addtionally the following turned letters representing printing errors has been replaced by appropriate correct characters:

```
LATIN SMALL LETTER TURNED A
LATIN SMALL LETTER TURNED E
LATIN SMALL LETTER TURNED M
```

The full list of the applied rules is available as the `onlyligatures4IMPACT.csv` file in the repository `https://bitbucket.org/jsbien/pol/`.

[5]Textel means a *text element*, cf. `http://bc.klf.uw.edu.pl/160/`

6 The search engine usage

Every page is a separate DjVu document, but the metadata `source` field contains the link to the appropriate page in a digital library.

The facsimile transcription is placed in the Poliqarp fields `base` intended for words lemmata, so to display it you have to select in the setting the option *Show lemmata* (in match, in context or both). To search e.g. for words which in facsimile transcription contain the MUFI character `LATIN ABBREVIATION SIGN SPACING BASE-LINE CAPITAL US` the field have to be referenced explictly in the query:

```
[base="\uf1a5"/x]
```

The textel transcription is placed in the Poliqarp fields `orth` intended for the actual spelling of words and other tokens called *segment*, so they are displayed without the need of any action on the user side. This field is also implicit in the query.

It strongly recommended to use in the queries the full power of regular expressions, especially POSIX equivalence class expressions, e.g.

```
"[[=a=]].*"
```

will match all the words (in the textel transcription) starting with

```
LATIN SMALL LETTER A
LATIN SMALL LETTER A WITH ACUTE
LATIN SMALL LETTER A WITH OGONEK
LATIN SMALL LETTER A WITH STROKE
LATIN SMALL LETTER A WITH DOT ABOVE
```

etc.

7 License

A Character histogram for facsimile transcription

This is the slightly hand-edited output of the unihistext program mentioned earlier.

The table lists both simple and composed characters. It contains:

1. Relative frequency as a percentage.
2. Absolute frequency.
3. The character code point as a hexadecimal digit or a sequence of code points.
4. The character itself, if available in the font used.

5. The name of the character or the sequence of names; the names of characters from Private Use Area are tagged as MUFI (Medieval Unicode Font Initiative) or Aletheia (a tool of the IMPACT project).

```
13.151  1061078                0x000020    SPACE
 7.061   569744                0x000069   i LATIN SMALL LETTER I
 6.290   507532                0x00006F   o LATIN SMALL LETTER O
 6.243   503688                0x000065   e LATIN SMALL LETTER E
 5.340   430815                0x000061   a LATIN SMALL LETTER A
 3.990   321930                0x00006E   n LATIN SMALL LETTER N
 3.885   313473                0x000079   y LATIN SMALL LETTER Y
 3.622   292230                0x000072   r LATIN SMALL LETTER R
 3.358   270957                0x00007A   z LATIN SMALL LETTER Z
 2.899   233927                0x000063   c LATIN SMALL LETTER C
 2.868   231427                0x000077   w LATIN SMALL LETTER W
 2.481   200188                0x000074   t LATIN SMALL LETTER T
 2.377   191804                0x00006D   m LATIN SMALL LETTER M
 2.280   183924                0x000064   d LATIN SMALL LETTER D
 2.173   175331                0x000075   u LATIN SMALL LETTER U
 2.157   174069                0x00006B   k LATIN SMALL LETTER K
 1.902   153429                0x00006C   l LATIN SMALL LETTER L
 1.796   144872                0x00002C   ,COMMA
 1.745   140800                0x00017F   ſ LATIN SMALL LETTER LONG S
 1.705   137590                0x000070   p LATIN SMALL LETTER P
 1.454   117343                0x00000A   LINE FEED (LF)
 1.317   106263                0x0000E1   á LATIN SMALL LETTER A WITH ACUTE
 1.201    96942                0x000062   b LATIN SMALL LETTER B
 1.176    94891                0x000142   ł LATIN SMALL LETTER L WITH STROKE
 1.123    90647                0x000067   g LATIN SMALL LETTER G
 0.975    78699                0x00002E   . FULL STOP
 0.910    73411                0x000068   h LATIN SMALL LETTER H
 0.692    55866                0x00EADA   ft LATIN SMALL LIGATURE LONG S DESCENDING T (MUFI)
 0.623    50280                0x00017C   ż LATIN SMALL LETTER Z WITH DOT ABOVE
 0.566    45666                0x002011   - NON-BREAKING HYPHEN
 0.514    41486                0x000105   ą LATIN SMALL LETTER A WITH OGONEK
 0.507    40923                0x000053   S LATIN CAPITAL LETTER S
 0.491    39605                0x000119   ę LATIN SMALL LETTER E WITH OGONEK
 0.481    38806                0x000050   P LATIN CAPITAL LETTER P
 0.460    37142                0x000073   s LATIN SMALL LETTER S
 0.452    36452                0x000041   A LATIN CAPITAL LETTER A
 0.404    32577                0x002C65   ⱥ LATIN SMALL LETTER A WITH STROKE
 0.391    31523                0x00004B   K LATIN CAPITAL LETTER K
 0.372    30020                0x00015B   ś LATIN SMALL LETTER S WITH ACUTE
 0.369    29748                0x00EBA2   ﮢ LATIN SMALL LIGATURE LONG S I (MUFI)
 0.349    28139                0x00004D   M LATIN CAPITAL LETTER M
 0.336    27111                0x000043   C LATIN CAPITAL LETTER C
 0.333    26895                0x000107   ć LATIN SMALL LETTER C WITH ACUTE
 0.328    26479                0x000049   I LATIN CAPITAL LETTER I
 0.309    24905                0x00004F   O LATIN CAPITAL LETTER O
 0.308    24856                0x000052   R LATIN CAPITAL LETTER R
 0.301    24286                0x000247   ɇ LATIN SMALL LETTER E WITH STROKE
 0.266    21471                0x00004E   N LATIN CAPITAL LETTER N
 0.258    20851                0x000054   T LATIN CAPITAL LETTER T
 0.255    20589                0x00002F   / SOLIDUS
 0.254    20481                0x000057   W LATIN CAPITAL LETTER W
 0.217    17506                0x000042   B LATIN CAPITAL LETTER B
 0.214    17301                0x000045   E LATIN CAPITAL LETTER E
 0.200    16153                0x00003A   : COLON
 0.194    15634                0x00005A   Z LATIN CAPITAL LETTER Z
 0.192    15531                0x000044   D LATIN CAPITAL LETTER D
 0.171    13825                0x00017A   ź LATIN SMALL LETTER Z WITH ACUTE
 0.169    13632                0x000047   G LATIN CAPITAL LETTER G
 0.166    13400                0x00004C   L LATIN CAPITAL LETTER L
 0.151    12165                0x000144   ń LATIN SMALL LETTER N WITH ACUTE
 0.149    12052                0x00003B   ; SEMICOLON
 0.142    11496                0x000076   v LATIN SMALL LETTER V
 0.135    10870                0x0000DF   ß LATIN SMALL LETTER SHARP S
 0.133    10722                0x000048   H LATIN CAPITAL LETTER H
 0.132    10622                0x000031   1 DIGIT ONE
 0.130    10526                0x00FFFD   ￿ REPLACEMENT CHARACTER
 0.117     9473                0x00004A   J LATIN CAPITAL LETTER J
 0.116     9364                0x000066   f LATIN SMALL LETTER F
 0.104     8379                0x000030   0 DIGIT ZERO
 0.102     8224                0x000055   U LATIN CAPITAL LETTER U
 0.100     8096                0x00F51E   ﴞ LATIN SMALL LIGATURE LONG S L WITH STROKE (Aletheia)
 0.093     7476                0x000046   F LATIN CAPITAL LETTER F
 0.091     7341                0x002E17   ⸗ DOUBLE OBLIQUE HYPHEN
 0.090     7278                0x000032   2 DIGIT TWO
 0.079     6358                0x000227   ȧ LATIN SMALL LETTER A WITH DOT ABOVE
 0.077     6240                0x000059   Y LATIN CAPITAL LETTER Y
 0.068     5481                0x0000E6   æ LATIN SMALL LETTER AE
 0.067     5396                0x000078   x LATIN SMALL LETTER X
 0.066     5350                0x000033   3 DIGIT THREE
 0.064     5158                0x0000E0   à LATIN SMALL LETTER A WITH GRAVE
 0.063     5083        0x000061 0x000301  á LATIN SMALL LETTER A, COMBINING ACUTE ACCENT
 0.062     5021                0x000034   4 DIGIT FOUR
 0.062     5020                0x000035   5 DIGIT FIVE
 0.061     4942                0x00017E   ž LATIN SMALL LETTER Z WITH CARON
```

```
0.054     4339                0x000036   6 DIGIT SIX
0.053     4294                0x000056   V LATIN CAPITAL LETTER V
0.047     3791                0x000058   X LATIN CAPITAL LETTER X
0.046     3672        0x00007A 0x000304   z̄ LATIN SMALL LETTER Z WITH COMBINING MACRON
0.042     3401                0x000037   7 DIGIT SEVEN
0.042     3395                0x00010B   ċ LATIN SMALL LETTER C WITH DOT ABOVE
0.040     3244                0x000026   & AMPERSAND
0.040     3227                0x00FB01   ﬁ LATIN SMALL LIGATURE FI
0.039     3159                0x00EEC5    LATIN SMALL LIGATURE CT (MUFI)
0.036     2932                0x000038   8 DIGIT EIGHT
0.036     2868                0x000071   q LATIN SMALL LETTER Q
0.033     2669                0x00EBA6    LATIN SMALL LIGATURE LONG S LONG S (MUFI)
0.032     2574                0x00006A   j LATIN SMALL LETTER J
0.032     2561                0x000039   9 DIGIT NINE
0.030     2459                0x000141   Ł LATIN CAPITAL LETTER L WITH STROKE
0.026     2063                0x00FB06    LATIN SMALL LIGATURE ST
0.025     1977                0x001E61   ṡ LATIN SMALL LETTER S WITH DOT ABOVE
0.019     1509                0x00003F   ? QUESTION MARK
0.018     1463                0x0000F3   ó LATIN SMALL LETTER O WITH ACUTE
0.016     1324                0x000028   ( LEFT PARENTHESIS
0.016     1279                0x000029   ) RIGHT PARENTHESIS
0.015     1224                0x00EBA7    LATIN SMALL LIGATURE LONG S LONG S I (MUFI)
0.014     1116                0x0000F2   ò LATIN SMALL LETTER O WITH GRAVE
0.013     1012                0x0000E8   è LATIN SMALL LETTER E WITH GRAVE
0.011      904                0x00007C   | VERTICAL LINE
0.009      761                0x0000C6   Æ LATIN CAPITAL LETTER AE
0.008      662                0x0000F4   ô LATIN SMALL LETTER O WITH CIRCUMFLEX
0.008      649        0x00007A 0x000309   ż LATIN SMALL LETTER Z, COMBINING HOOK ABOVE
0.008      627                0x0000E9   é LATIN SMALL LETTER E WITH ACUTE
0.008      606                0x000051   Q LATIN CAPITAL LETTER Q
0.007      585                0x00FB00   ﬀ LATIN SMALL LIGATURE FF
0.006      487                0x00FB02   ﬂ LATIN SMALL LIGATURE FL
0.005      397        0x000061 0x000300   à LATIN SMALL LETTER A, COMBINING GRAVE ACCENT
0.005      368                0x000246   Ɇ LATIN CAPITAL LETTER E WITH STROKE
0.004      346                0x0000E2   â LATIN SMALL LETTER A WITH CIRCUMFLEX
0.004      308                0x000021   ! EXCLAMATION MARK
0.004      296                0x000118   Ę LATIN CAPITAL LETTER E WITH OGONEK
0.003      269                0x00023A   Ⱥ LATIN CAPITAL LETTER A WITH STROKE
0.003      258                0x0000EE   î LATIN SMALL LETTER I WITH CIRCUMFLEX
0.003      255                0x00FB03   ﬃ LATIN SMALL LIGATURE FFI
0.003      253                0x00EBA3    LATIN SMALL LIGATURE LONG S L (MUFI)
0.003      223                0x000104   Ą LATIN CAPITAL LETTER A WITH OGONEK
0.002      199        0x000071 0x000301   q́ LATIN SMALL LETTER Q, COMBINING ACUTE ACCENT
0.002      190                0x00005D   ] RIGHT SQUARE BRACKET
0.002      185                0x001E45   ṅ LATIN SMALL LETTER N WITH DOT ABOVE
0.002      185                0x0000F9   ù LATIN SMALL LETTER U WITH GRAVE
0.002      183                0x0000FB   û LATIN SMALL LETTER U WITH CIRCUMFLEX
0.002      180                0x00016B   ū LATIN SMALL LETTER U WITH MACRON
0.002      177                0x00F1AC    LATIN ABBREVIATION SIGN SEMICOLON (MUFI)
0.002      166        0x00007A 0x000303   z̃ LATIN SMALL LETTER Z, COMBINING TILDE
0.002      166                0x00002D   -HYPHEN-MINUS
0.002      151                0x00005B   [ LEFT SQUARE BRACKET
0.002      139        0x00006E 0x000304   n̄ LATIN SMALL LETTER N, COMBINING MACRON
0.002      129                0x00017B   Ż LATIN CAPITAL LETTER Z WITH DOT ABOVE
0.001      119        0x000067 0x000366   g LATIN SMALL LETTER G, COMBINING LATIN SMALL LETTER O
0.001      116                0x0000B0   ° DEGREE SIGN
0.001      111                0x000113   ē LATIN SMALL LETTER E WITH MACRON
0.001       93                0x00E8BA    LATIN SMALL LETTER V WITH SHORT SLASH (MUFI)
0.001       91                0x000153   œ LATIN SMALL LIGATURE OE
0.001       90                0x00005C   \ REVERSE SOLIDUS
0.001       82                0x001E83   ẃ LATIN SMALL LETTER W WITH ACUTE
0.001       81                0x000101   ā LATIN SMALL LETTER A WITH MACRON
0.001       76                0x00E5DC    LATIN SMALL LETTER N WITH MEDIUM HIGH MACRON ABOVE (MUFI)
0.001       71                0x0000ED   í LATIN SMALL LETTER I WITH ACUTE
0.001       70        0x00006F 0x000300   ò LATIN SMALL LETTER O, COMBINING GRAVE ACCENT
0.001       70        0x000065 0x000300   è LATIN SMALL LETTER E, COMBINING GRAVE ACCENT
0.001       63                0x0000A7   § SECTION SIGN
0.001       62                0x000027   'APOSTROPHE
0.001       60                0x0000EC   ì LATIN SMALL LETTER I WITH GRAVE
0.001       48                0x000117   ė LATIN SMALL LETTER E WITH DOT ABOVE
0.001       47                0x00014D   ō LATIN SMALL LETTER O WITH MACRON
0.001       41                0x0000F1   ñ LATIN SMALL LETTER N WITH TILDE
0.000       38                0x00002A   * ASTERISK
0.000       36                0x0000F5   õ LATIN SMALL LETTER O WITH TILDE
0.000       36                0x000133   ij LATIN SMALL LIGATURE IJ
0.000       34                0x00F1A6    LATIN ABBREVIATION SIGN SPACING BASE-LINE US (MUFI)
0.000       33                0x001E87   ẇ LATIN SMALL LETTER W WITH DOT ABOVE
0.000       32                0x00015A   Ś LATIN CAPITAL LETTER S WITH ACUTE
0.000       30        0x000065 0x000301   é LATIN SMALL LETTER E, COMBINING ACUTE ACCENT
0.000       29                0x00FB04   ﬄ LATIN SMALL LIGATURE FFL
0.000       29        0x00006D 0x000303   m̃ LATIN SMALL LETTER M, COMBINING TILDE
0.000       29                0x0000EA   ê LATIN SMALL LETTER E WITH CIRCUMFLEX
0.000       27                0x00261E   ☞ WHITE RIGHT POINTING INDEX
0.000       27                0x00022F   ȯ LATIN SMALL LETTER O WITH DOT ABOVE
0.000       27                0x0000FA   ú LATIN SMALL LETTER U WITH ACUTE
0.000       24         0x00E8BF 0x000301    LATIN SMALL LETTER Q LIGATED WITH FINAL ET (MUFI), COMBINING ACUTE ACCENT
0.000       23                0x00E5B8    LATIN SMALL LETTER M WITH MEDIUM HIGH MACRON ABOVE (MUFI)
0.000       22                0x000169   ũ LATIN SMALL LETTER U WITH TILDE
0.000       21                0x0000EF   ï LATIN SMALL LETTER I WITH DIAERESIS
0.000       19                0x0000E3   ã LATIN SMALL LETTER A WITH TILDE
0.000       19        0x00006F 0x000301   ó LATIN SMALL LETTER O, COMBINING ACUTE ACCENT
```

0.000	18		0x001EBD	ẽ LATIN SMALL LETTER E WITH TILDE
0.000	16		0x00010D	č LATIN SMALL LETTER C WITH CARON
0.000	16		0x00211F	℟ RESPONSE
0.000	15		0x001E8F	ẏ LATIN SMALL LETTER Y WITH DOT ABOVE
0.000	15		0x0000B7	· MIDDLE DOT
0.000	13		0x0000E7	ç LATIN SMALL LETTER C WITH CEDILLA
0.000	12	0x00006E	0x000309	n̉ LATIN SMALL LETTER N, COMBINING HOOK ABOVE
0.000	11		0x00E8BF	ﷺ LATIN SMALL LETTER Q LIGATED WITH FINAL ET (MUFI)
0.000	11		0x001E59	ṙ LATIN SMALL LETTER R WITH DOT ABOVE
0.000	10		0x001EF9	ỹ LATIN SMALL LETTER Y WITH TILDE
0.000	10		0x000179	Ź LATIN CAPITAL LETTER Z WITH ACUTE
0.000	8		0x000233	ȳ LATIN SMALL LETTER Y WITH MACRON
0.000	8		0x0000EB	ë LATIN SMALL LETTER E WITH DIAERESIS
0.000	8		0x0000FD	ý LATIN SMALL LETTER Y WITH ACUTE
0.000	8		0x001E3F	ḿ LATIN SMALL LETTER M WITH ACUTE
0.000	7		0x0000FC	ü LATIN SMALL LETTER U WITH DIAERESIS
0.000	6		0x000106	Ć LATIN CAPITAL LETTER C WITH ACUTE
0.000	6	0x0000A0	0x000366	ͦ NO-BREAK SPACE, COMBINING LATIN SMALL LETTER O
0.000	6		0x001E41	ṁ LATIN SMALL LETTER M WITH DOT ABOVE
0.000	6		0x00005F	_ LOW LINE
0.000	5		0x002010	‐ HYPHEN
0.000	5		0x001E31	ḱ LATIN SMALL LETTER K WITH ACUTE
0.000	5	0x000063	0x000304	c̄ LATIN SMALL LETTER C, COMBINING MACRON
0.000	5		0x001E82	Ẃ LATIN CAPITAL LETTER W WITH ACUTE
0.000	4	0x000075	0x000300	ù LATIN SMALL LETTER U, COMBINING GRAVE ACCENT
0.000	4	0x000073	0x000304	s̄ LATIN SMALL LETTER S, COMBINING MACRON
0.000	4		0x0000E4	ä LATIN SMALL LETTER A WITH DIAERESIS
0.000	4	0x00006F	0x000302	ô LATIN SMALL LETTER O, COMBINING CIRCUMFLEX ACCENT
0.000	4		0x00026F	ɯ LATIN SMALL LETTER TURNED M
0.000	4		0x00A770	ꝰ MODIFIER LETTER US
0.000	4		0x00A75F	ꝟ LATIN SMALL LETTER V WITH DIAGONAL STROKE
0.000	4		0x0000F6	ö LATIN SMALL LETTER O WITH DIAERESIS
0.000	3		0x0000D3	Ó LATIN CAPITAL LETTER O WITH ACUTE
0.000	3		0x001E57	ṗ LATIN SMALL LETTER P WITH DOT ABOVE
0.000	3		0x0001F5	ǵ LATIN SMALL LETTER G WITH ACUTE
0.000	3	0x000075	0x000302	û LATIN SMALL LETTER U, COMBINING CIRCUMFLEX ACCENT
0.000	3		0x00F4F9	ﴹ LATIN SMALL LIGATURE LL (MUFI)
0.000	3		0x00201E	„ DOUBLE LOW-9 QUOTATION MARK
0.000	3		0x0000C9	É LATIN CAPITAL LETTER E WITH ACUTE
0.000	3		0x001E9E	ẞ LATIN CAPITAL LETTER SHARP S
0.000	3		0x001ECD	ọ LATIN SMALL LETTER O WITH DOT BELOW
0.000	2		0x002018	‘ LEFT SINGLE QUOTATION MARK
0.000	2		0x00F1A5	 LATIN ABBREVIATION SIGN SPACING BASE-LINE CAPITAL US (MUFI)
0.000	2		0x0001DD	ə LATIN SMALL LETTER TURNED E
0.000	2		0x001E60	Ṡ LATIN CAPITAL LETTER S WITH DOT ABOVE
0.000	2	0x000063	0x000300	c̀ LATIN SMALL LETTER C, COMBINING GRAVE ACCENT
0.000	2		0x0001F9	ǹ LATIN SMALL LETTER N WITH GRAVE
0.000	2		0x000115	ĕ LATIN SMALL LETTER E WITH BREVE
0.000	2	0x000075	0x000304	ū LATIN SMALL LETTER U, COMBINING MACRON
0.000	2		0x000121	ġ LATIN SMALL LETTER G WITH DOT ABOVE
0.000	2		0x001EF3	ỳ LATIN SMALL LETTER Y WITH GRAVE
0.000	2		0x00A753	ꝓ LATIN SMALL LETTER P WITH FLOURISH
0.000	2		0x00A751	ꝑ LATIN SMALL LETTER P WITH STROKE THROUGH DESCENDER
0.000	2	0x000061	0x000302	â LATIN SMALL LETTER A, COMBINING CIRCUMFLEX ACCENT
0.000	2		0x00016D	ŭ LATIN SMALL LETTER U WITH BREVE
0.000	2		0x001E89	ẉ LATIN SMALL LETTER W WITH DOT BELOW
0.000	2		0x002720	✠ MALTESE CROSS
0.000	2		0x0000C1	Á LATIN CAPITAL LETTER A WITH ACUTE
0.000	2	0x000065	0x000306	ĕ LATIN SMALL LETTER E, COMBINING BREVE
0.000	2	0x000077	0x000306	w̆ LATIN SMALL LETTER W, COMBINING BREVE
0.000	1	0x000072	0x000304	r̄ LATIN SMALL LETTER R, COMBINING MACRON
0.000	1		0x00010A	Ċ LATIN CAPITAL LETTER C WITH DOT ABOVE
0.000	1		0x0000AC	¬ NOT SIGN
0.000	1	0x000072	0x00030A	r̊ LATIN SMALL LETTER R, COMBINING RING ABOVE
0.000	1	0x000075	0x000301	ú LATIN SMALL LETTER U, COMBINING ACUTE ACCENT
0.000	1		0x001EA3	ả LATIN SMALL LETTER A WITH HOOK ABOVE
0.000	1		0x000122	Ģ LATIN CAPITAL LETTER G WITH CEDILLA
0.000	1		0x00012E	Į LATIN CAPITAL LETTER I WITH OGONEK
0.000	1		0x00A76B	ꝫ LATIN SMALL LETTER ET
0.000	1		0x001ECB	ị LATIN SMALL LETTER I WITH DOT BELOW
0.000	1	0x000072	0x000300	r̀ LATIN SMALL LETTER R, COMBINING GRAVE ACCENT
0.000	1		0x0000FF	ÿ LATIN SMALL LETTER Y WITH DIAERESIS
0.000	1	0x000071	0x000304	q̄ LATIN SMALL LETTER Q, COMBINING MACRON
0.000	1		0x002013	– EN DASH
0.000	1		0x00201B	‛ SINGLE HIGH-REVERSED-9 QUOTATION MARK
0.000	1		0x000023	# NUMBER SIGN
0.000	1		0x00201C	“ LEFT DOUBLE QUOTATION MARK
0.000	1		0x0000C0	À LATIN CAPITAL LETTER A WITH GRAVE
0.000	1		0x0000C8	È LATIN CAPITAL LETTER E WITH GRAVE
0.000	1		0x0001CE	ǎ LATIN SMALL LETTER A WITH CARON
0.000	1		0x001E55	ṕ LATIN SMALL LETTER P WITH ACUTE
0.000	1		0x00017D	Ž LATIN CAPITAL LETTER Z WITH CARON
0.000	1		0x001E81	ẁ LATIN SMALL LETTER W WITH GRAVE
0.000	1		0x000103	ă LATIN SMALL LETTER A WITH BREVE
0.000	1		0x00261C	☜ WHITE LEFT POINTING INDEX
0.000	1		0x00022E	Ȯ LATIN CAPITAL LETTER O WITH DOT ABOVE
0.000	1		0x001EA1	ạ LATIN SMALL LETTER A WITH DOT BELOW
0.000	1	0x000079	0x000300	ỳ LATIN SMALL LETTER Y, COMBINING GRAVE ACCENT
0.000	1	0x000079	0x000304	ȳ LATIN SMALL LETTER Y, COMBINING MACRON
0.000	1		0x00015F	ş LATIN SMALL LETTER S WITH CEDILLA
0.000	1		0x001EB9	ẹ LATIN SMALL LETTER E WITH DOT BELOW

```
0.000        1                0x0000CD   Í LATIN CAPITAL LETTER I WITH ACUTE
0.000        1                0x00E665   □ LATIN SMALL LETTER P WITH MACRON (MUFI)
0.000        1        0x000131 0x000301   í LATIN SMALL LETTER DOTLESS I, COMBINING ACUTE ACCENT
0.000        1                0x000250   ɐ LATIN SMALL LETTER TURNED A
0.000        1                0x000394   Δ GREEK CAPITAL LETTER DELTA
0.000        1        0x000077 0x000304   w̄ LATIN SMALL LETTER W, COMBINING MACRON
0.000        1                0x0000D2   Ò LATIN CAPITAL LETTER O WITH GRAVE
0.000        1                0x0000F8   ø LATIN SMALL LETTER O WITH STROKE
0.000        1        0x000065 0x000303   ẽ LATIN SMALL LETTER E, COMBINING TILDE
```